W. SWARTOUT, J. GRATCH, R. HILL, E. HOVY, R. LINDHEIM,
S. MARSELLA, J. RICKEL, D. TRAUM

SIMULATION MEETS HOLLYWOOD:

Integrating Graphics, Sound, Story and Character for Immersive Simulation

Abstract. The Institute for Creative Technologies was created at the University of Southern California with the goal of bringing together researchers in simulation technology to collaborate with people from the entertainment industry. The idea was that much more compelling simulations could be developed if researchers who understood state-of-the-art simulation technology worked together with writers and directors who knew how to create compelling stories and characters.

This paper presents our first major effort to realize that vision, the Mission Rehearsal Exercise Project, which confronts a soldier trainee with the kinds of dilemmas he might reasonably encounter in a peacekeeping operation. The trainee is immersed in a synthetic world and interacts with virtual humans: artificially intelligent and graphically embodied conversational agents that understand and generate natural language, reason about world events and respond appropriately to the trainee's actions or commands. This project is an ambitious exercise in integration, both in the sense of integrating technology with entertainment industry content, but also in that we have also joined a number of component technologies that have not been integrated before. This integration has not only raised new research issues, but it has also suggested some new approaches to difficult problems. In this paper we describe the Mission Rehearsal Exercise system and the insights gained through this large-scale integration.

1. INTRODUCTION

For many researchers, software integration is often regarded as a kind of necessary evil – something that must be done to make sure that all the research components of a large system fit together and interoperate properly – but not something that is likely to contribute new research insights or suggest new solutions. Our work on constructing virtual humans to interact with people in virtual environments has involved large-scale integration of a number of software technologies that support the simulation of human behaviors, ranging from speech recognition and dialogue management through task reasoning, gesture generation and emotion modeling. In addition, because we use the virtual humans in training simulations, the characters behave in the context of a scenario, so another aspect of integration has been to bring together story content with virtual human behavior.

In integrating these various components and content, we have been surprised to find that the conventional wisdom about integration does not hold: the integration process has raised new research issues and at the same time has suggested new approaches to long-standing issues. This paper describes how that has taken place and our discoveries. We begin with a brief description of the background behind our work in training and the approach we have taken to improving training. We then

T.H.E. Editor(s) (ed.), Book title, 1—6.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Simulation Meets Hollywood: Integrating Graphics, Sound, Story and Character for Immersive Simulation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Institute for Creative Technologies,13274 Fiji Way,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**25**	

describe the technology components we have developed, the system architecture we use, and we conclude with the insights we have gained from the integration process.

1.1. Background

How can training simulations be made more effective? An important insight in answering that question is to recognize that effective training depends both on the technology that is used to present the material and the content of the material itself. The Institute for Creative Technologies (ICT) was created at the University of Southern California with the goal of bringing together researchers in simulation technology to collaborate with people from the entertainment industry. The idea was that if those who understood how to create high resolution graphics, immersive sound, and believable virtual humans worked together with those who understood how to create compelling stories and characters a synergy would emerge that would allow them to create much more compelling simulation experiences.

Although the ICT has only been in existence for a short time, we are already beginning to see some of the results of this collaboration. These are reflected both in the kinds of projects that the ICT takes on and the approach that we take to implementing systems. While most military simulations involve simulating a vehicle such as a tank, an airplane or a helicopter, ICT's simulations put trainees into a human-oriented simulation, where they interact with real and virtual (computer-generated) humans. While scenarios in most military simulations tend to proceed in a straightforward fashion, our scenarios engage the trainee with plot twists, turns and surprises, much like one might find in a good Hollywood script. In constructing our simulations we have used a hybrid approach, mixing different techniques and technologies to produce the best overall effect. In that way, we are following Hollywood film production techniques where what appears as a single seamless scene in film may actually be the result of integrating a large number of disparate elements produced using filmed live action, computer generated imagery, and models.

One of the ICT's projects that illustrates these ideas well is the Mission Rehearsal Exercise (MRE) project. Since the end of the cold war, the kinds of operations that the US military is involved with has expanded greatly. The need for peacekeeping and nation-building operations has grown, and humanitarian efforts such as disaster relief are common. One of the hallmarks of these operations is that they frequently involve close interactions between the military and the local civilian populace. To function effectively and avoid misunderstandings that could have unintended consequences, it is important that soldiers understand the customs, norms, habits and taboos of the local population and they need to be exposed to the thorny dilemmas and decisions that may await them.

The Mission Rehearsal Exercise system, shown in Figure 1, is designed to provide that kind of experience in simulation, before trainees encounter it in reality. Presented on a 30 foot by 8 foot curved screen, the MRE system places the trainee in a location. The trainee interacts with life-sized virtual humans that can play the role

Figure 1: The Mission Rehearsal Exercise System, showing from the left, the platoon sergeant, the injured boy and his mother, a medic, and a crowd.

of local civilians, friendly forces and hostile forces. A 10.2 sound system (10 channels of audio, 2 subwoofer channels) enhances the immersive effect.

The scenario we are currently using is situated in a small town in Bosnia. It opens with a lieutenant (the trainee) in his Humvee. Over the radio, he gets orders to proceed to a rendezvous point to meet up with his soldiers to plan a mission to assist in quelling a civil disturbance. When he arrives at the rendezvous point, he discovers a surprise. One of his platoon's Humvees has been involved in an accident with a civilian car. There's a small boy on the ground with serious injuries, a frantic mother, and a crowd is starting to form. A TV camera crew shows up and starts taping. What should the lieutenant do? Should he stop and render aid? Or should he continue on with his mission? Depending on decisions he makes, different outcomes will occur. The initial version of the Mission Rehearsal Exercise system was first shown in September, 2000. Since then, the MRE project has been actively engaged in research to improve the MRE system and make it more interactive.

2. MRE ARCHITECTURE

The MRE architecture, illustrated in Figure 2, supports the flexible integration of a number of components, including visualization components (such as graphics and audio processing), interface components (such as voice input) and behavioral components (such as virtual humans and the scenario manager). Components are linked through a messaging and notification service (the communication bus). Here we consider the communication services, graphics and animation, audio processing, and some of the behavior modeling. The details of the virtual human architecture are discussed in the following section.

2.1. Communication Services

Components in the MRE system communicate primarily through a common communications bus, implemented through a notification and messaging service called Elvin that enables efficient inter-process and cross-platform communication

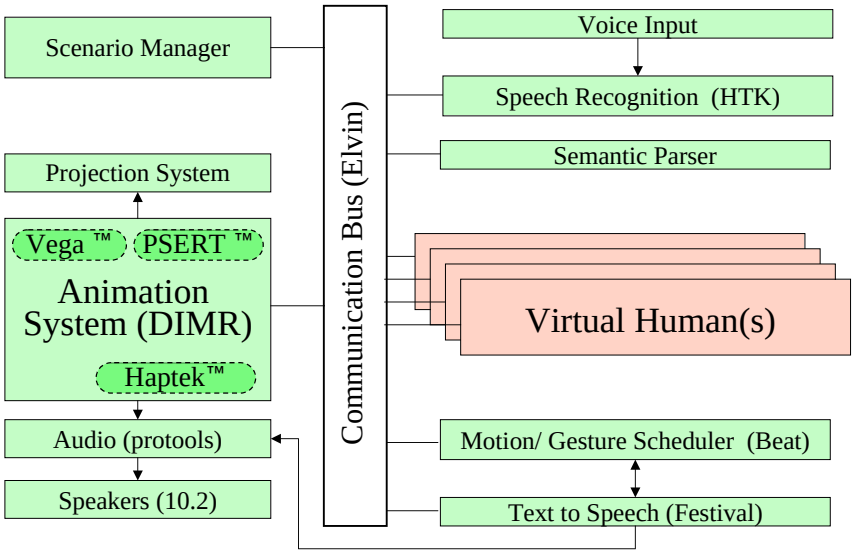


Figure 2: MRE System Architecture

(elvin.dstc.edu.au). Elvin uses a client-server architecture and allows messages to be filtered based on their content to reduce network load. Components send all their messages to the server and messages are routed to individual components if they have registered interest in the specific message type. Message content is formatted as simple text strings or XML, facilitating the easy creation of new message types or formats.

Two communication pathways bypass Elvin for efficiency purposes. There is a dedicated communication link between the Animation System and the audio system to mitigate latencies and, similarly, there is also a dedicated link between the text-to-speech engine and the character gesture manager (BEAT).

2.2. Graphics and Animation

The graphics and animation system, DIMR, provides a set of core services for visualizing activities in the virtual world. DIMR uses two commercial products, Vega™ and PeopleShop™, to animate the virtual world. Vega™ renders the environment and the special effects. The environment includes the buildings, roads, trees, vehicles, and so on, while the special effects include explosions and the dynamic motion of objects like cars and helicopters. The PeopleShop™ Embedded Runtime System (PSERT) is integrated with Vega™ and provides the animation of the characters' bodies. A 3D model of a Balkan village was developed to fit the types of scenarios we had in mind. Texture mapped surfaces were applied to the buildings, vehicles, and characters to give them a more authentic look and feel.

Boston Dynamics Incorporated (BDI), the developers of PeopleShop™, extended their virtual character bodies in several ways to suit our needs. First, they integrated expressive faces (developed by Haptek Incorporated) to support lip synchronization and facial expressions. Second, while the basic PeopleShop™ software primarily supports dynamic sequencing of primitive motion fragments, BDI combined their motion-capture approach with procedural animation to provide more flexibility, primarily in the areas of gaze and arm gestures (Marsella, Gratch & Rickel, 2003). Finally, we wanted more variety when it came to the character bodies, so BDI created a suite of new character bodies and behaviors. The new character bodies included a Balkan woman (to play the mother), a child, a man holding a TV news camera, civilian characters for the crowd, an Army medic, and an Army sergeant.

2.3. Audio Processing

In our current scenario the scene begins with the lieutenant driving up to the village in an Army vehicle known as a Humvee. As the vehicle drives into town and turns a corner, our view out the front windshield and side windows allows us to see the road, buildings, and trees. We perceive the bumps in the road as a jiggle in the scene, and the vehicle appears to change velocity as the gears are shifted. While the visual aspects of the scene give the viewer a sense of being there in that village, the audio system provides a critical dimension to the experience. The distinctive roar of the Humvee's diesel engine, the creaks, rattles, and bumps from the bouncy ride, and the post-ignition knock when the engine shuts off are all synchronized with visual effects. When the lieutenant steps out of the Humvee, one can immediately hear the murmur of a crowd of people speaking in Serbo-Croatian, gathered near the accident site. When the medevac helicopter flies overhead the room literally vibrates with the sound of a Blackhawk helicopter.

To address the problem of matching picture with sound spatially, a novel multi-channel audio system was developed (Kyriakakis, 1998). This system uses 10 channels of audio and 2 subwoofer channels. Speakers are not only arranged in a radial pattern around the participant, similar to conventional surround sound, but in addition, there is a vertical displacement between speakers. This allows sound to be spatialized in both horizontal and vertical dimensions, creating, in effect, a hemisphere of sound around the audience. This means that the sound of a helicopter flyover will be perceived as coming from overhead, making the sonic experience much more convincing.

2.4. Behavior Modeling

Some of the behaviors in the MRE system are autonomous, but others are specified in advance based on the story that a writer develops. Generally, the behaviors of the major characters in the simulation are autonomous, while physical events (e.g. explosions) and minor characters are scripted. The scenario manager component triggers scripted behaviors to shape the experience for the trainee and create the

dilemmas he must solve. Scripted behaviors can be triggered either autonomously, through a set of simple production rules, or by a human exercise controller.

There are currently two classes of agents playing the character roles in the MRE system: scripted characters and virtual humans controlled by AI. The scripted (minor) characters come packaged with PeopleShop™. They can be scripted to perform specific actions, such as running along a pre-specified path or securing a perimeter, and this behavior can be triggered by the scenario manager or a virtual human. The scripted characters do not perceive anything in the world — their behaviors are generated by playing motion capture sequences. Virtual humans, the major characters, are autonomous and their structure is described next.

3. VIRTUAL HUMANS

Our virtual humans build on prior work in the areas of embodied conversational agents (Cassell, Sullivan, Prevost, & Churchill, 2000) and animated pedagogical agents (Johnson, Rickel, & Lester, 2000), but they integrate a broader set of capabilities than any prior work. For the types of training scenarios we are targeting, the virtual humans must integrate three broad influences on their behavior: they must perceive and act in a 3D virtual world, they must engage in face-to-face spoken dialogues with people and other virtual humans in such worlds, and they must exhibit human-like emotions. Classic work on virtual humans in the computer graphics community focused on perception and action in 3D worlds (Badler, Phillips, & Webber, 1993; Thalmann, 1993), but largely ignored dialogue and emotions. Several systems have carefully modeled the interplay between speech and nonverbal behavior in face-to-face dialogue (Cassell, Bickmore, Campbell, Vilhjálmsson, & Yan, 2000; Cassell et al., 1994; Pelachaud, Badler, & Steedman, 1996) but these virtual humans did not include emotions and could not participate in physical tasks in 3D worlds. Some work has begun to explore the integration of conversational capabilities with emotions (Lester, Towns, Callaway, Voerman, & FitzGerald, 2000; Marsella, Johnson, & LaBore, 2000; Poggi & Pelachaud, 2000), but still does not address physical tasks in 3D worlds. Likewise, our prior work on Steve addressed the issues of integrating face-to-face dialogue with collaboration on physical tasks in a 3D virtual world (Rickel & Johnson, 1999a, 1999b, 2000), but Steve did not include emotions and had far less sophisticated dialogue capabilities than our current virtual humans. The tight integration of all these capabilities is one of the most novel aspects of our current work.

The virtual humans, which include the sergeant, medic, and mother in the scenario, are implemented in Soar, a general architecture for building intelligent agents (Newell, 1990) and build on the earlier Steve system. As such, their behavior is not scripted; rather, it is driven by a set of general, domain-independent capabilities discussed below. The virtual humans perceive events in the simulation, reason about the tasks they are performing, and they control the bodies and faces of the PeopleShop™ characters to which they have been assigned. They send messages to one another, to the character bodies, and to the audio system via the Communications Bus shown in Figure 3.

3.1. Virtual Human Architecture

In order for virtual humans to collaborate with people and each other in scenarios like the peacekeeping mission with a sufficient illusion of human-like behavior to keep human users engaged, they must include a wide variety of capabilities, such as perception, planning, spoken dialogue, and emotions. Our research objectives are to advance the state of the art in each of these areas, but also to explore their integration into a single agent architecture. Thus, we desired a flexible architecture for our virtual humans that would allow us to easily experiment with the connections between the individual components.

A blackboard architecture, in which individual components have access to the intermediate and final results of other components, provides such flexibility. The alternative, in which each module would explicitly pass specific information to other components, would require constant revision as we made progress understanding the interdependencies among components. In contrast, a blackboard architecture would make all intermediate and final results of individual components available by default, so the designers of each component could make use of such results as they proved useful.

For our integrated architecture, we chose Soar, because it allows each component to be implemented with production rules that read from and write to a common working memory, which acts as the desired blackboard. Soar further breaks computation into a sequence of intermediate *operators* that are proposed in parallel

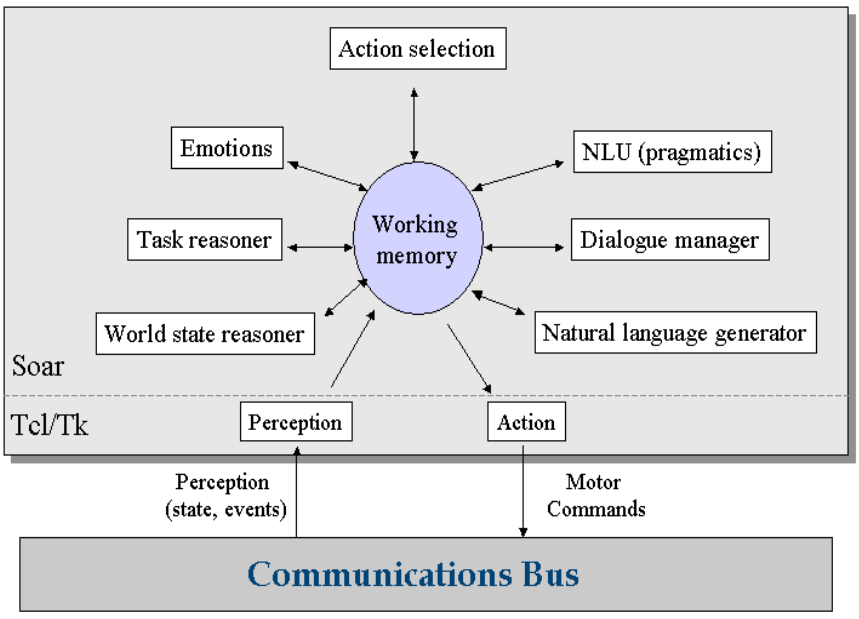


Figure 3: Virtual Human architecture

but selected sequentially via an arbitration mechanism. This allows for tight interleaving of operators from individual components and flexible control over their priority.

All components of the virtual humans are implemented in Soar, with several exceptions: speech recognition, natural language understanding (syntactic and semantic analysis), synchronization of verbal and nonverbal components of output utterances, and speech synthesis. It was less practical to implement these four components in Soar because each was built on top of existing software that would have been difficult to reimplement.

3.2. Task Representation and Reasoning

To collaborate with humans and other synthetic teammates, virtual humans need to understand how past events, present circumstances, and future possibilities impact team tasks and goals. For example, the platoon sergeant agent must be able to brief the trainee on past events that led to the accident and must reason how the victim's current injuries impact the platoon's mission. More generally, agents must understand task goals and how to assess whether they are currently satisfied, the actions that can achieve them, how the team must coordinate the selection and execution of those actions, and how to adapt execution to unexpected events. To provide this understanding, our agents use domain-independent reasoning algorithms operating over a general, declarative representation of team tasks, and this representation is used to encode their domain-specific task knowledge for a given training scenario (or class of scenarios).

The agent maintains an explicit representation of past, present and future task-related information in Soar's working memory. This representation extends our earlier work on virtual humans for team training (Rickel & Johnson, 2002) and includes three components: the task description, a causal history, and the current world description.

The task description includes of a set of steps, each of which is either a primitive action (e.g., a physical or sensing action in the virtual world) or an abstract action which must itself be further decomposed. Abstract actions give tasks a hierarchical structure. There may be ordering constraints among the steps, which define a partial order. Interdependencies among steps in the task description or causal history are represented as a set of causal links and threat relations (McAllester & Rosenblitt, 1991). Each causal link specifies that an effect of a step in the task could achieve a particular goal that is a precondition for another step in the task (or for termination of the task). For example, in our military domain there is an action of marking a landing zone with smoke, which achieves the goal of allowing a helicopter pilot to visually identify the landing zone, which in turn is a precondition for landing it. Threat relations specify that an effect of a step could threaten a causal link by unachieving the goal before it is needed. For example, extinguishing the smoke before the helicopter arrives threatens the helicopter's ability to land.

The causal history maintains a sequence of executed steps (including unexpected and non-task events), interdependencies between past steps (e.g., causal links), as well as interdependencies between past steps and future steps in the task description.

In addition to understanding the structure of tasks, agents must understand the roles of each team member. Each task step is associated with the team member that is responsible for performing it (Rickel & Johnson, 2002). We have also extended our representation to include an optional association of each task step with the teammate who has authority over its execution; that is, the teammate responsible for a task step cannot perform it until authorization is given by the specified teammate with authority (Traum et al., 2003). This extension to the representation was required to model the hierarchical organizational structure of some teams, such as in the military.

Given a top-level abstract task for the team to accomplish, each agent independently uses its task knowledge to construct a complete task model. Starting with the task description for the top-level task, the agent recursively expands any abstract step with its task description, until the agent has a fully decomposed, hierarchical task model. Agents may or may not be given identical task knowledge, and so may or may not construct identical task models; this can be used to model teammates with partial or erroneous knowledge.

An agent's task model represents its understanding of the task in general, independent of the current scenario conditions. To guide execution of the task and robustly handle unexpected events that require adaptive execution or replanning, agents use a partial-order planning algorithm over the task model; the algorithm is described in detail in (Rickel & Johnson, 1999a), and its application to reasoning about team tasks is detailed in (Rickel & Johnson, 2002). The task model specifies all the steps that might be required to complete the task; it can be viewed as a worst-case plan. Agents continually monitor the state of the virtual world via messages from the simulator (Rickel & Johnson, 1999a) that are filtered to reflect perceptual limitations (Rickel et al., 2002). These perceptions will allow the agents to update their representations of the status of goals in the task model as being satisfied, unsatisfied, or unknown if they cannot currently perceive the state of the goal. The planning algorithm works backwards through the causal links in the task model to identify goals that are currently desired and task steps that are currently intended to establish those desired goals. Just as the status of a goal can be satisfied, unsatisfied, or unknown, the planning algorithm marks the “desired” property of goals and the “intended” property of steps as true, false, or unknown. The result of this planning algorithm specifies how the agent privately believes that the team can collectively complete the task, with some causal links specifying the interdependencies among team members' actions. Agents continually revise this private plan as the scenario unfolds.

A key aspect of collaborative planning is negotiating about alternative ways to achieve team goals (Traum et al., 2003). To support such negotiation, we have extended our earlier representation so that task models support reasoning about alternative, mutually exclusive courses of action (recipes) for achieving tasks, and we have added mechanisms for evaluating the relative strengths and weaknesses of different alternatives. These courses of action are self-contained hierarchical tasks

in the sense defined above, and subject to the same dynamic task reasoning. For example, one might evacuate someone to a hospital by using either a medevac helicopter or an ambulance. Depending on the circumstances, only one option might be possible (e.g., the medevac may be unavailable or the injuries may be too severe for an ambulance), but if both are valid options, they must be ranked through some reasoned analysis of their relative costs and benefits.

3.3. Natural Language Dialogue

In many ways, our natural language processing components and architecture mirror fairly traditional dialogue systems. There is a speech recognizer, semantic parser, dialogue manager, NL generator, and speech synthesizer. However, the challenges of the MRE project, including integration within an immersive story environment as well as with the other virtual human components required innovations in most areas. Here we briefly describe the natural language processing components and capabilities; we will return later to some of the specific innovations motivated by this integration.

The Speech recognizer was built using the Hidden Markov Model Toolkit (http://htk.eng.cam.ac.uk/) currently employing a limited domain finite-state language model with a several hundred word vocabulary and using about 70 phrases, and with locally trained acoustic models (Wang & Narayanan, 2002). Output is currently the single best interpretation, sent as Elvin messages, as well as indications of when the user starts and stops speaking, to manage gaze control and turn-taking behavior of agents.

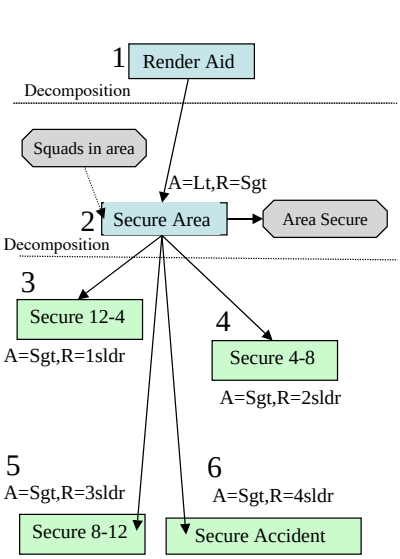
Speech recognition output is processed by the semantic parser module, which produces a semantic representation of the utterances. The parser uses a hybrid between finite-state transducers and statistical processing to produce a best-guess at semantic information from the input word stream (Feng 2003). In cases in which imperfect input is given, it will robustly produce representations which may possibly be incomplete or partially incorrect. The module will provide addressee information (if vocatives were present), sentence mood, and semantic information corresponding to states and actions related to the task model. See (Traum, 2003) for more details about the semantic representation.

The SOAR-module for each agent receives the output of the speech recognizer and semantic parser. This information is then matched against the agent's internal representation of the context, including the actions and states in the task model, current expectations, and focus to determine a set of candidate interpretations. Some of these interpretations may be underspecified, due to impoverished input, or over-specified in cases of incorrect input (either an out of domain utterance by the user, or an error in the speech recognizer or semantic parser). In some cases, underspecified elements can be filled in with reference to the agent's knowledge; if not, the representation is left underspecified and processing continues. The dialogue component of the SOAR agent also produces a set of dialogue act interpretations of the utterance. Some of these are traditional speech acts (e.g., assert, request, info-request) with content being the semantic interpretation, while others represent other

levels of action that have been performed, such as turn-taking, grounding, and negotiation. See (Traum & Rickel, 2002) for details on the levels of dialogue acts.

Dialogue management follows the approach of the TRINDI Project (Larsson & Traum, 2000), and specifically the EDIS system (Matheson, Poesio, & Traum, 2000). Dialogue acts are used to update an *Information State* that is also used as context for other aspects of agent reasoning. SOAR is actually very similar to the TrindiKit software used by EDIS, so it was straightforward to adapt the prior dialogue update rules into the SOAR agent. More on the aspects of information state can be found in (Traum & Rickel, 2002). Decisions of how to act in dialogue are tightly coupled with other action selection decisions in the agent. The agent can choose to speak, choose to listen, choose to act related to a task, etc. Aspects of the information state provide motivations to speak, including answering questions, negotiating with respect to a request or order, giving feedback of understanding (acknowledgements, repairs, and repair requests), and making suggestions and issuing orders, when appropriate according to the task model.

Once a decision is made to speak, there are several phases involved in the language production process. First is the *content selection* phase, in which the agent reasons about how best to achieve the output goal. Examples are which assertion to make to answer a pending question, or how to respond to a negotiation proposal. Once the content has been selected, next there is a *sentence planning* phase, deciding the best way to convey this content. The output of this phase is a case frame structure that specifies the content and some aspects of the form of each utterance. Next, *realization* proceeds in two passes. In the first pass, each noun phrase unit is realized as a variety of alternatives. As described later, units with the most appropriate emotional connotations are selected. In the second pass, variations



Focus=1
Lt: U11"secure the area"
Commited(lt,2), 2 authorized,Obl(sgt,U11)
Sgt: U12"yes sir"
Committed(sgt,2), Push(2,focus)
Goal7:Announce(2,{1sldr,2sldr,3sldr,4sldr})
Goal8: Start-conversation(sgt, ,{1sldr,2sldr,…},2)
Goal8 -> Sgt: U21 "Squad leaders listen up!"
Goal7 -> U22 "I want 360 degree security"
Push(3, focus)
Goal9:authorize 3
Goal9 -> u23"1st squad take 12-4"
Committed(sgt,3), 3 authorized
Pop(3), Push(4)
Goal10: authorize 4
Goal10 -> u24"2nd squad take 4-8"
Committed(sgt,4), 4 authorized
Pop(4)
…
A10: Squads move
A10: grounds U21-U26,…
ends conversation about 2, realizes 2

Figure 4: Sample task model and dialogue interaction

of the sentence itself are realized, using the selected noun phrases, and then similarly ranked for connotations. Finally, the sentence that maximizes the inclusion of semantic content and the expression of desired emotional connotations is selected. This final sentence is then augmented with communicative gestures and sent to the synthesizer and rendering modules to produce the speech. Meanwhile, messages are sent to other agents, letting them know what the agent is saying. More details on the generation component can be found in (Fleischman & Hovy, 2002; Traum, Fleischman, & Hovy, 2003). The speech synthesizer uses Festival and Festvox, with locally developed unit-selection limited-domain voices to provide the emotional expressiveness needed to maintain immersiveness (Johnson et al., 2002).

Figure 4 shows a brief example of how dialogue behavior is integrated with task reasoning. The left side of the figure shows a small fragment of the task model: part of the "Render aid" task involves securing the assembly area, which requires that the squads are in the area; it has a decomposition involving actions of various squads, and has the effect that the area is secure. The figure also shows which agents are responsible (R) for seeing that an action is performed (either doing it themselves or acting as team leader making sure the subtasks are carried out), and which agents have authority (A) to have the action performed. With reference to this piece of the task model, consider the dialogue fragment on the right. Initially the focus is on the render aid task. When the lieutenant issues the command to secure the area (utterance U11), the sergeant recognizes the command as referring to a subaction of Render Aid in the current task model (Task 2). As a direct effect of the lieutenant issuing a command to perform this task, the lieutenant becomes committed to the task, the sergeant has an obligation to perform the task, and the task becomes authorized. Because the sergeant already agrees that this is an appropriate next step, he is able to accept it with utterance U12, which also commits him to perform the action. The sergeant then pushes this task into his task model focus and begins execution. In this case, because it is a team task requiring actions of other teammates, the sergeant, as team leader, must announce the task to the other team members. Thus, the system forms a communicative goal to make this announcement. Before the sergeant can issue this announcement, he must make sure he has the squad leaders' attention and has them engaged in conversation. He forms a goal to open a new conversation so that he can produce the announcement. Then his focus can turn to the individual tasks for each squad leader. As each one enters the sergeant's focus, he issues the command that commits the sergeant and authorizes the troops to carry it out. When the sergeant observes the troops move into action, he can infer that they have understood his order and adopted his plan. When the task completes, the conversation between sergeant and squad leaders finishes and the sergeant turns his attention to other matters.

3.4. Emotion

Our work on modeling emotion is motivated by the Cognitive Appraisal theory of emotion. Cognitive Appraisal is a psychological theory of emotion that emphasizes the relationship between emotion and cognition (Lazarus, 1991). The theory posits

two basic processes: appraisal and coping. Appraisal generates emotion by assessing the person-environment relationship (did an event facilitate or inhibit the agent's goals; who deserves blame or credit). Coping is the process of dealing with emotion, either by acting externally on the world (problem-focused coping), or by acting internally to change beliefs or attention (emotion-focused coping). Coping and appraisal interact and unfold over time, modeling the temporal character of emotion noted by several emotion researchers (Lazarus, 1991; Scherer, 1984): an agent may "feel" distress for an event (appraisal), which motivates the shifting of blame (coping), which leads to anger (re-appraisal).

In re-casting this theory as a computational model, we have tied appraisals and coping to the explicit representation of past, present, and future task-related information in Soar's working memory, discussed above. This representation has several advantages for modeling emotion. It makes a clean separation between domain-specific knowledge (e.g., specific action definitions, probabilities and utilities) from the domain-independent mechanisms that operate on these representations. It acts as a blackboard architecture, simplifying communication between appraisal and coping to other mechanisms (like planning) that operate on the interpretation. It facilitates reasoning about blame and indirect consequences of action (e.g., a threat to a sub-goal might be distressing, not because the sub-goal is intrinsically important, but because it facilitates a larger goal). It provides a uniform representation of past and future actions (this action caused an effect which I can use to achieve that goal). Finally, it facilitates reasoning about different agents' perspectives (I think this outcome is good but I believe you think it is bad).

Our approach to appraisal assesses the agent-environment relationship via features of this explicit task representation (Gratch, 2000). Speaking loosely, we treat appraisal as a set of feature detectors that map features of this representation into appraisal variables that characterize the consequences of an event from the agent's perspective. These variables include the desirability of those consequences, the likelihood of them occurring, who deserves credit or blame and a measure of the agent's ability to alter those consequences. The result of this feature detection is one or more data structures, called appraisal frames, which characterize the agent's emotional reactions to an event. Thus, the belief that another agent has caused an undesirable outcome leads to distress and possibly anger.

Our computational model of coping -- as described in (Marsella & Gratch, 2002) -- similarly exploits the task representation to uncover which features led to the appraised emotion, and what potential there may be for altering these features. In essence, coping is the inverse of appraisal. To discharge a strong emotion about some situation, one obvious strategy is to change one or more of the factors that contributed to the emotion. Coping operates on the same representations as the appraisals, the agent's beliefs, goals and plans, but in reverse, seeking to make a change, directly or indirectly, that would have the desired impact on appraisal. Coping could impact the agent's beliefs about the situation, such as the importance of a threatened goal, the likelihood of the threat, responsibility for the threat, etc. Further, the agent might form intentions to change external factors, for example, by performing some action that removes the threat. Indeed, our coping strategies can involve a combination of such approaches. This mirrors how coping processes are

understood to operate in human behavior whereby people may employ a mix of problem-focused coping and emotion-focused coping to deal with stress.

Coping behavior is focused by those Soar operators that update the task representation and thus helps to reveal the emotional significance and can inform the prioritization of those operations. At any point in time, the virtual humans have many different emotions corresponding to multiple features of the task representation. To perform in the virtual environment an agent must understand and generate speech, generate and repair plans and direct its sensors to perceive activities in the environment. All of these operations reference or modify the agent's interpretation of past, present or future task-related information. For example, perception updates beliefs. Each time one of these operations accesses an element of the task representation it activates any emotional appraisals associated with the element. These emotions associated with the object are made available as "concerns" for the coping process.

Whereas there has been prior work in computational models of appraisal, there has been little prior work in modeling the myriad ways that people cope with emotions. And yet coping behavior is a key aspect of human behavior. People employ a rich set of coping strategies and different individuals tend to adopt stable and characteristic "coping styles" that are correlated with personality type. Our work is building a library of these strategies and uses personality-inspired preference rules to model consistent differences in style across different agents. For example, our virtual humans may take preemptive action to circumvent a stressful factor, they may choose to shift blame to another agent or they may behaviorally disengage from attempts to achieve a goal that is being thwarted or threatened.

3.5. Body Movements

Internally, the virtual humans are continually perceiving the events surrounding them, understanding utterances, updating their beliefs, formulating and revising plans, generating emotional appraisals, and choosing actions. Our goal is to manifest the rich dynamics of this cognitive and emotional inner state through each character's external behavior using the same verbal and nonverbal cues that people use to understand one another. The key challenge is the range of behaviors that must be seamlessly integrated: each character's body movements must reflect its awareness of events in the virtual world, its physical actions, the myriad of nonverbal signals that accompany speech during social interactions (e.g., gaze shifts, head movements, and gestures), and its emotional reactions.

Since gaze indicates a character's focus of attention, it is a key element in any model of outward behavior, and must be closely synchronized to the character's inner thoughts. Prior work on gaze in virtual humans has considered either task-related gaze (Chopra-Khullar & Badler, 2001) or social gaze (Cassell et al., 1994) but has not produced an integrated model of the two. Our gaze model is driven by our cognitive model, which interleaves task-related behaviors, social behaviors, and attention capture. Task-related behaviors (e.g., checking the status of a goal or monitoring for an expected effect or action) trigger a corresponding gaze shift, as

does attention capture (e.g., hearing a new sound in the environment). Gaze during social interactions is driven by the dialogue state and the state of the virtual human's own processing, including gaze at an interlocutor who is speaking, gaze aversion during utterance planning (to claim or hold the turn), gaze at an addressee when speaking, and gaze when expecting someone to speak. This tight integration of gaze behaviors to our underlying cognitive model ensures that the outward attention of the virtual humans is synchronized with their inner thoughts.

Body movements are also critical for conveying emotional changes, including facial expressions, gestures, posture, gaze and head movements (Marsella, Gratch, & Rickel, 2001; Marsella, Gratch & Rickel 2003). In humans, these behaviors are signals and as such they can be used intentionally by an individual to inform or deceive but can also unintentionally reveal information about the individual's internal emotional state. Thus a person's behavior may express anger because they feel it or because they want others to think they feel it or for both reasons. Prior work on emotional expression in virtual humans has focused on either the intentional emotional expression or as a window on internal emotional state (Neal Reilly, 1996). Our work attempts to integrate these aspects by tying expressive behavior to coping behavior. As noted earlier, emotional changes in the virtual human unfold as a consequence of Soar operators updating the task representation. These operators provide a focus for emotional processes, invoking coping strategies to address the resulting emotions which in turn leads to expressive behaviors. This focus on operators both centers emotional expression on the agent's current internal cognitive processing but also allows coping to alter the relation of the expression to those internal cognitive processes. Thus, when making amends, our virtual humans might freely express their true appraisal-based feelings of guilt and concern, for example through facial expressions, gestures, posture, gaze and head movements. However, when shifting responsibility, it might suppress an initial expression of guilt and rather express anger at the character they are blaming, to reflect a more calculated attempt to persuade others.

Finally, a wide range of body movements are typically closely linked to speech, movements that emphasize, augment and even supplant components of the spoken linguistic information. Consistent with this close relation, this nonverbal behavior, which can include hand-arm gestures, head movements and postural shifts, is typically synchronized in time with the speech. Realizing this synchronization faces the challenge that we do not have an incremental model of speech production. Such a model would allow us to tie nonverbal behaviors to speech production operations much like the gaze and coping behaviors are tied to cognitive operations. Rather, our approach is to plan the utterance out and annotate it with nonverbal behavior. The annotated utterance is then passed to a text-to-speech generation system that schedules both the verbal and nonverbal behavior, using the BEAT system (Cassell, Vilhjálmsson, & Bickmore, 2001). This approach is similar to the work of Cassell et al. (Cassell et al., 1994). Our work differs in the structure passed to the gesture annotation process, in order to capture the myriad ways that the nonverbal behavior can relate to the spoken dialog and the internal state of the virtual human. Specifically, while both systems pass the syntactic, semantic and pragmatic structure of the utterance, we additionally pass the emotional appraisal and coping

information associated with the components of the utterance. The gesture annotation process uses this information to annotate the utterance with gestures, head movements, eyebrow lifts and eyeflashes.

4. THE ROLE OF STORY

The AI-driven virtual humans in MRE are autonomous, as of course is the human trainee. However, there is an overall scenario or story that sets the context and shapes the experience for the trainee. Since there are certain pedagogical goals we want to achieve for the trainee, we feel that it is necessary to provide structure and guidance to the experience he has. If he is allowed to wander aimlessly through the simulation, he may never encounter the decision-making dilemmas we want him to experience. Thus story is critical to the training experience, and the quality of the story can determine the degree of engagement the trainee feels.

The function and value of story is often misunderstood and misrepresented. Story is not simply a progression of events. Passive (or traditional) storytelling is an integration of elements, harmoniously interacting to create a seamless, involving experience. Interactive storytelling adds further complexity to the task.

In its best form, passive storytelling is predictably unpredictable. It is predictive because the ultimate outcome of the story is usually ordained, and the audience expects the outcome. Most good stories contain elements of parable and morality. Some typical predictable outcomes are: good triumphs over evil; justice prevails; hard work is rewarded; immoral acts are punished. Stories use the element of verisimilitude to create a coherent world, where this message can be delivered to the reader/listener/viewer.

While stories are usually relegated to entertainment, the concept of storytelling can be interwoven into the fabric of the story to provide the pedagogical goals of learning. In this form story becomes an essential element of simulation training.

The unpredictable element of a story is critical but infrequently understood. It is said that there are only five to seven basic stories. Yet, there have been hundreds of thousands of stories told, if not millions. And we, the reader/listener/viewer feel the experience is new with each story. How is that so? It is the application of unpredictability to the equation. While the outcome of the story is pre-ordained, the way in which the story is told and the actual events that occur within the story can vary almost infinitely. The story need not even be told linearly (consider the film "Momento.") It is these sudden shifts in events, character behaviors, and environment that engender the reader/listener/viewer involvement. Creating these unpredictable elements is the essential function of the writer, and the brilliance by which they accomplish this end separates the hack from the award winner.

Interactive storytelling compounds the task of the writer. The predictable element of the story remains much the same. The pedagogic/morality aspect is relatively unchanged. And the story still progresses forward in time from beginning to end. But interactive storytelling vastly increases the complexity of the unpredictable element. The participant in the interactive experience gains a measure of control of events, and the writer must use a different construct to keep the values

of the story secure and still provide a compelling experience. This is accomplished by devising a support structure for the story, much in the way that a road surface enables a car to move easily forward. The support structure also provides the triggers for the unpredictable elements that allow the writer to maintain control of the immersive experience.

This story support structure is also critical to the integration of story with other interactive elements in the simulation. As we will discuss in the next section, the integration of story with the technical aspects of the simulation can synergistically enhance the simulation experience.

5. PUTTING IT ALL TOGETHER: THE VALUE OF INTEGRATION

We have described the major technical components of the Mission Rehearsal Exercise system and the story content that the MRE scenario is based on. As we pointed out in the introduction, software integration is necessary to make sure that all the various pieces in a system work together properly, but one usually expects that the real research takes place in building the individual components. One doesn't expect to learn much from integration (expect perhaps to find that some components don't interface properly). However, in integrating the Mission Rehearsal Exercise system, we have been surprised: we have uncovered new research issues and some new approaches to existing problems have been suggested. In this section we outline some of the things we learned as we brought all the pieces together.

5.1. Dialogue in rich social environments with multiple characters

As we have argued above, a good story involves compelling characters with intriguing interrelationships. Supporting such an environment imposes new demands on natural language processing. In particular, the Bosnian scenario of MRE, with a cast of many characters occupying various roles in a rich social fabric, is quite different from the usual case of natural language dialogue with a single human and single computer system interacting. While some aspects of dialogue as social interaction had already been addressed in previous work (e.g, discourse obligations in (Traum & Allen, 1994)), many new issues needed to be addressed to allow agents to understand and engage in this sort of dialogue. These issues include

- Is the intended addressee paying attention?
- Is he already engaged in conversation?
- How will hearers recognize who is the addressee?
- How are vocatives and gaze as well as context reasoning used to help this process?
- How are multiple, interleaved, conversations managed (e.g., talking face to face with one character while on the radio to another)?

These issues have implications for agents in both understanding and producing communications, and for representing the dialogue state. Furthermore, there are differences depending on whether the conversation is between virtual humans or

between the human trainee and a virtual human, because more limited information is available in the second case.

We have begun to address these issues in several ways. First, the dialogue model has been extended so that who is being addressed is captured as well as the content to be conveyed. Second, we have introduced conventions for marking the start and termination of a conversation with an agent. A conversation begins by addressing the character either by name or by his role. For example the lieutenant might give the sergeant an order by saying: “Sergeant, send first squad to Celic!” Once a conversation has been started, it is assumed to continue until it is terminated, either by the purpose having been fulfilled (for a short task-specific conversation like securing the area), or by an explicit closing (e.g., "out" on the radio).

For conversations between the human trainee and the virtual humans we rely on these conventions to determine who is addressing whom. For conversations between virtual humans, the problem of determining who is being addressed is easier, because it is all represented internally. However, the virtual humans use the same reasoning methods when talking among themselves as they use for interacting with the trainee so their behavior is consistent. We feel this is an important constraint to achieve consistency in interface behavior. See (Traum & Rickel, 2002) for additional details on our work in this area. In the future, we would like to make use of head-tracking data to determine who the trainee is looking at when he speaks. This is an additional source of information that could be used to determine whom he is addressing.

We are just beginning to scratch the surface in this area, and we hope to see more sophisticated techniques emerge as research progresses. But the surprising thing to us is that this area of inquiry has received so little attention from the computational linguistic community, yet it is clearly basic to multi-person interactions. It points out to us the value of large-scale integration that has forced us to confront this new research issue.

5.2. The Pervasive Effect of Emotion

In humans, emotion has a broad effect on behavior. It affects how we speak, how we gesture, our posture, and even how we reason. And, of course, emotion is indispensable for creating good story and compelling characters. In integrating emotion into our virtual humans, we have found that we need to deal with a similarly broad range of issues. Models of emotion can both affect the behavior of other components of the virtual human, and they can provide additional knowledge that the system can use in reasoning. Below we give an example of each.

5.2.1 Emotionally Appropriate Natural Language Generation

A big challenge for Natural Language Generation in MRE is the generation of emotionally appropriate language, which expresses both the desired information and the desired emotional attitude towards that information. Each expressive variant casts an emotional shade on each representational item it contains (for example, the

phrase governed by the verb "ram" as in "They rammed into us, sir" casts the subject in a negative and the object in a positive light). Prior work on the generation of variation expressions, such as (Bateman & Paris, 1989; Hovy, 1990), uses quite simplistic emotional models of the speaker and hearer. In general, these systems simply had to choose among a small set of phrases, and within the phrase from a small set of lexical fillers for certain positions of the phrase, where each alternative phrase and lexical item was pre-annotated with an affective value such as *good* or *bad*.

The presence in MRE of an emotion model provides a considerably finer-grain level of control, enabling principled realization decisions over a far more nuanced set of expressive alternatives. Given many representational items, a rich set of emotional values potentially holding for them, and numerous phrases, each with its own combination of positive and negative fields, the problem was to design a system that can reliably and quickly find the optimal phrasing without dropping content. (Of course, in some cases no perfect solution may exist. The best way to say "we crashed into them" may be "they were bumped", but it omits part of the material to be conveyed.) Emotion-based realization involves a potentially expensive process of casting representation items into phrase positions with appropriate connotations, where different positions may have different strengths, and making sure that the phrases themselves cover the material to be conveyed. To compute shades of connotation more accurately and quickly, we created a vector space in which we can represent the desired attitudes of the speaker (as specified by the emotion model) as well as the overall emotional value of each candidate expression (whether noun phrase or whole sentence). Using a standard Euclidean distance measure we can then determine which variant expression most closely matches the desired effect.

After realization has produced all variants for a given input, and determined their distances from the emotion model's desired value, the ranking algorithm then combines the distance scores with a score reflecting how much of the input content was in fact realized in the output. The overall winner is selected and passed along for speech synthesis. More details on the NLG Module can be found in (Fleischman & Hovy, 2002).

5.2.2 Using Emotion to Determine Linguistic Focus

In natural language, we often refer to things in imprecise ways. To correctly interpret such referents in a natural language utterance, one needs to understand what is in linguistic focus. Loosely speaking, one needs to understand what is the main subject of discussion. For example, when the lieutenant trainee arrives at the accident scene in the MRE scenario, he might ask the sergeant, "What happened here?" In principle many things have happened: the lieutenant just drove up, the soldiers assembled at the meeting point, an accident occurred, a crowd formed, and so forth. The sergeant could talk about any one of these and be factually correct, but he would sound quite silly if he responded: "Well, you just drove up, sir." The expected response is for the sergeant to talk about the accident. To produce an

appropriate response the sergeant needs to understand that the accident is in linguistic focus.

A number of heuristics have been developed to model linguistic focus. One such heuristic is based on the idea of *recency*. It holds that the entity that is in linguistic focus is whatever was most recently discussed, or occurred most recently. In this case, recency doesn't work, since the lieutenant opens the conversation with his question and several things have happened subsequent to the accident.

However, people are often focused most strongly on the things that upset them emotionally, which suggests an emotion-based heuristic for determining linguistic focus. Because we have modeled the sergeant's emotions in MRE the linguistic routines have access to the fact that he is upset about the accident and they can use that information in determining linguistic focus, allowing the sergeant to give the most appropriate answer and describe the accident and how it occurred.

5.3. Integration of Story with Virtual (and Real) Humans

A well-constructed story can play a critical role in enhancing a trainee's experience. We have identified several ways in which this can occur.

First, because the technology is still immature, interactivity with virtual humans driven by artificial intelligence is constrained. AI characters have limited intelligence and range of activities. The story support structure must be aware of these limitations and other factors within the simulation environment and respond accordingly. For example, because the BDI characters do not support collision detection, characters cannot touch each another or be in close proximity; a small error in positioning would make them simply move through one another, destroying the illusion. For similar reasons, it is very difficult to animate the soldiers climbing into the Humvee. The writer must build such constraints into the story support system. In the case of the Humvee the easy solution was to have the AI characters enter from the far side of the vehicle, so that the viewer would not be aware that they did not actually climb into it.

Second, the story support structure can use the element of unpredictability to distract the viewing audience from the flaws in the system. Motion picture and television people do this often. If they do not want the viewer to focus attention to a particular portion of the screen, they use distraction as a tool. For example, where budget constraints have made them use a painted backdrop instead of an actual location, the creative people create activity to draw the eye to that portion of the screen desired. And, because the viewer is willing to suspend disbelief, he or she does not focus upon it and accepts the patently false image of the painted backdrop as real. This use of story distraction is a powerful tool for the MRE simulation. Returning again to the difficulty of animating the soldiers getting into their vehicles, we have found that sometimes the solution suggested above of having the soldiers enter on the side of the vehicle that is out of sight does not work due to constraints in the layout of the scene. However, if the problematic action occurs in the background, and a distracting action takes place in the foreground, the problem will probably not be noticed. For example, at one point in the scenario, a number of

soldiers leave in their Humvees. Although the animation of the soldiers entering the vehicles is awkward, it occurs in the background. Furthermore, at the same time, the mother character's emotion model causes her to become very upset and start gesturing and shouting excitedly in the foreground, because she feels the soldiers are abandoning her and her son. This foreground action advances the story and at the same time distracts from the awkward animation, so viewers tend not to notice the flaws.

Third, when we consider the range of technologies that we are trying to integrate in MRE we realize making it all work is an ambitious goal. Indeed, we believe that if we were to try to construct virtual humans that could function in the real world and provide a wide range of capabilities such as speech recognition, natural language understanding and generation, emotion modeling and body animation, the task would be too hard for the current state of technology: the range of situations that the system would have to deal with would be too great. But we are not trying to build virtual humans that operate in the real world. Instead, we are building an artificial world *that we control via a story line* and introducing real people into it. The story provides a very strong context both from a rational and emotional perspective that limits the possible responses that the human trainee will make. This works because people are predictable in their responses. If a hundred people from the same culture are put into the same situation, they won't respond in a hundred different ways. Instead a handful of responses will cover the range of responses. A story, by providing that strong context, very much limits the range of responses the system must handle, which in turn limits the breadth and range of knowledge that must be programmed into the virtual characters. The limited testing we have performed so far has confirmed this hypothesis although additional testing is needed.

This integration of the predictable and unpredictable elements of storytelling in MRE demonstrates the critical role they play in creating an immersive training simulation.

6. STATUS

An initial version of the MRE system described in this paper has been implemented and applied to the peacekeeping training scenario described earlier. The system allows the trainee, playing the role of the lieutenant, to interact freely (through speech) with the three virtual humans (sergeant, medic, and mother). The trainee's primary interaction is with the sergeant, who is the main source of information about what happened and advice about how to proceed. The trainee takes action in the virtual world through commands to the sergeant, who in turn commands the squads. Ultimately, the experience terminates with one of four possible endings, depending on the trainee's actions. However, unlike interactive narrative models based on an explicit branching structure, the system does not force the trainee through a predetermined sequence of decision points, each with a limited set of options; the trainee's interactions with the characters is unconstrained and limited only by the characters' understanding and capabilities.

The understanding and capabilities of the virtual humans is limited by the coverage of their spoken dialogue models and their models of the domain tasks. The sergeant's speech recognizer currently has a vocabulary of a few hundred words, with a grammar allowing recognition of 16000 distinct utterances. His natural language understanding module can currently produce semantic representation frames for all of these sentences as well as providing (sometimes partial) results for different or ill-formed input. His natural language generation module currently expresses all communicative goals formed by the dialog module, modulating some of them for affective appropriateness. His speech synthesis module currently has a vocabulary of over 1000 words. The sergeant's domain task knowledge, which is the most complex among all the virtual humans in the scenario, includes about 40 tasks, and about 150 properties of the world. While the tasks represent the full range of actions that the sergeant can understand and carry out, his ability to talk about these tasks and properties (e.g., answer questions and give advice) is broad, limited only by the coverage of the spoken dialogue modules as described above.

Despite its complexity, real-time performance of the system is good, although we are continuing to improve latencies. Given an utterance by the user, a virtual human typically responds within 3 seconds, including speech recognition, natural language understanding, updating dialogue and emotional states, choosing how to respond, natural language generation, planning the voice output and accompanying gestures and visemes, and finally producing the speech. As is typical of humans, the virtual humans are producing communicative behaviors throughout this time delay, including averting gaze from the user during the utterance planning phases to indicate that they are formulating a response (Kendon 1967).

We have tested the system with a variety of users acting as trainees. Early sessions were useful for system debugging, but since these trainees lacked the military background required to understand the appropriate actions in situations such as our peacekeeping scenario, sessions were not useful for formal evaluations. In general, trainees with some knowledge of the scenario were often successful in using the system but were undoubtedly biased by their knowledge, and those without such knowledge often failed because they had little idea of how to proceed in such situations. We have just begun testing the system with trainees who have more appropriate military backgrounds, and we expect to report our results in a forthcoming paper.

7. SUMMARY

Integration is a kind of two-edged sword. Making a large number of components work together requires a significant effort in developing a system architecture and the interfaces between the components. But as we have tried to illustrate in this paper, integration can also open up new vistas for research and it can enable new solutions to difficult problems. To us, this suggests that integration needs to be thought of as an integral part of the research process, rather than something that is done once all the research is complete.

8. ACKNOWLEDGEMENTS

The authors would like to thank A. Crane, W. Crane, J. Deweese, J. Douglas, D. Feng, M. Fleischman, W.L. Johnson, Y. J. Kim, S. Kwak, C. Kyriakakis, C. LaBore, A. Marshall, D. Miraglia, B. Moore, J. Morie, M. Murguia, S. Narayanan, P. O'Neal, D. Ravichandran, M. Raibert, M. Thiébaux, L. Tuch, M. Veal, and R. Whitney for their hard work and enthusiasm that contributed greatly to this research.

This paper was developed with funds of the United States Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the United States Department of the Army.

9. REFERENCES

Badler, N. I., Phillips, C. B., & Webber, B. L. (1993). *Simulating Humans*. New York: Oxford University Press.

Bateman, J. A., & Paris, C. L. (1989). *Phrasing a Text in Terms the User can Understand.* Paper presented at the 11th International Joint Conference on Artificial Intelligence, Detroit, MI.

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., & Yan, H. (2000). Human conversation as a system framework: Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied Conversational Agents* (pp. 29-63). Boston: MIT Press.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., et al. (1994). *Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents.* Paper presented at the ACM SIGGRAPH, Reading, MA.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.). (2000). *Embodied Confersational Agents*. Cambridge, MA: MIT Press.

Cassell, J., Vilhjálmsson, H., & Bickmore, T. (2001). *BEAT: The Behavior Expressive Animation Toolkit.* Paper presented at the SIGGRAPH, Los Angeles, CA.

Chopra-Khullar, S., & Badler, N. (2001). Where to Look? Automating Attending Behaviors of Virtual Human Characters. *Autonomous Agents and Multi-Agent Systems, 4*(1-2), 9-23.

Feng, D (2003) Cooperative Model Based Language Understanding in Dialogue. To be presented at the NAACL/HLT Student Research Workshop, Edmonton.

Fleischman, M., & Hovy, E. (2002). Emotional variation in speech-based natural language generation. Paper presented at the International Natural Language Generation Conference, Arden House, NY.

Gratch, J. (2000). Émile: Marshalling Passions in Training and Education. Paper presented at the Fourth International Conference on Intelligent Agents, Barcelona, Spain.

Hovy, E. H. (1990). Pragmatics and Natural Language Generation. *Artificial Intelligence, 43(2)*, 153-198.

Johnson, W. L., Narayanan, S., Whitney, R., Das, R., Bulut, M., & LaBore, C. (2002). *Limited Domain Synthesis of Expressive Military Speech for Animated Characters.* Paper presented at the IEEE Workshop on Speech Synthesis, Santa Monica, CA.

Johnson, W. L., Rickel, J., & Lester, J. C. (2000). Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of AI in Education, 11*, 47-78.

Kendon, A. (1967). Some Functions of Gaze Direction in Two-Person Conversation. *Acta Psychologica, 26*, 1–47.

Kyriakakis, C. (1998). Fundamental and Technological Limitations of Immersive Audio Systems. *Proceedings of the IEEE, 86(5)*, 941-951.

Larsson, S., & Traum, D. (2000). Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering, 6*, 323-340.

Lazarus, R. (1991). *Emotion and Adaptation*. NY: Oxford University Press.

Lester, J. C., Towns, S. G., Callaway, C. B., Voerman, J. L., & FitzGerald, P. J. (2000). Deictic and Emotive Communication in Animated Pedagogical Agents. In J. Cassell, S. Prevost, J. Sullivan & E. Churchill (Eds.), *Embodied Conversational Agents* (pp. 123-154). Cambridge: MIT Press.

Marsella, S., & Gratch, J. (2002). *A Step Toward Irrationality: Usign Emotion to Change Belief.* Paper presented at the First International Joint Conference on Autonomous Agents and Multiagent Systems, Bologna, Italy.

Marsella, S., Gratch, J., & Rickel, J. (2001). *The Effect of Affect: Modeling the Impact of Emotional State on the Behavior of Interactive Virtual Humans.* Paper presented at the Agents 2001 Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents, Montreal, Canada.

Marsella, S., Gratch, J., & Rickel, J. (2003) Expressive Behaviors for Virtual Worlds, in *Like-like Characters. Tools, Affective Functions and Applications*, Helmut Prendinger and Mitsuru Ishizuka (eds.), Springer-Verlag (in press).

Marsella, S., Johnson, W. L., & LaBore, C. (2000). *Interactive Pedagogical Drama.* Paper presented at the Fourth International Conference on Autonomous Agents, Montreal, Canada.

Matheson, C., Poesio, M., & Traum, D. (2000). *Modeling Grounding and Discourse Obligations Using Update Rules.* Paper presented at the First Conference of the North American Chapter of the Association for Computational Linguistics.

McAllester, D., & Rosenblitt, D. (1991). *Systematic Nonlinear Planning.* Paper presented at the Ninth National Conference on Artificial Intelligence, Menlo Park, CA.

Neal Reilly, W. S. (1996). *Believable Social and Emotional Agents* Ph.D Thesis No. CMU-CS-96-138. Pittsburgh, PA: Carnegie Mellon University.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Pelachaud, C., Badler, N. I., & Steedman, M. (1996). Generating Facial Expressions for Speech. *Cognitive Science, 20*(1).

Poggi, I., & Pelachaud, C. (2000). Emotional Meaning and Expression in Performative Faces. In A. Paiva (Ed.), *Affective Interactions: Towards a New Generation of Computer Interfaces*. Berlin: Springer-Verlag.

Rickel, J., & Johnson, W. L. (1999a). Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control. *Applied Artificial Intelligence, 13*, 343-382.

Rickel, J., & Johnson, W. L. (1999b). *Virtual Humans for Team Training in Virtual Reality.* Paper presented at the Ninth International Conference on Artificial Intelligence in Education.

Rickel, J., & Johnson, W. L. (2000). Task-Oriented Collaboration with Embodied Agents in Virtual Worlds. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied Conversational Agents*. Boston: MIT Press.

Rickel, J., & Johnson, W. L. (2002). Extending Virtual Humans to Support Team Training. In G. Lakemayer & B. Nebel (Eds.), *Exploring Artificial Intelligence in the New Millenium* (pp. 217-238). San Francisco: Morgan Kaufmann.

Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D., & Swartout, W. (2002). Toward a New Generation of Virtual Huans for Interactive Experiences. *IEEE Intelligent Systems, July/August,* 32-38.

Scherer, K. (1984). On the Nature and Function of Emotion: A Component Process Approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 293-317).

Thalmann, D. (1993). Human Modeling and Animation. In *Eurographics '93 State-of-the-Art Reports*.

Traum, D. (2003). *Semantics and Pragmatics of Questions and Answers for Dialogue Agents.* Paper presented at the Fifth International Workshop on Computational Semantics, Tilburg.

Traum, D., & Allen, J. F. (1994). *Discourse Obligations in Dialogue Processing.* Paper presented at the 32nd Annual Meeting of the Association for Computational Linguistics.

Traum, D., Fleischman, M., & Hovy, E. (2003). *NL Generation for Virtual Humans in a Complex Social Environment.* paper presented at the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue.

Traum, D., & Rickel, J. (2002). *Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds.* Paper presented at the First International Conference on Autonomous Agents and Multi-agent Systems, Bologna, Italy.

Traum, D., Rickel, J., Gratch, J. and Marsella, S.. "Negotiation over Tasks in Hybrid Human-Agent Teams for Simulation-Based Training", to appear in Proceedings of Autonomous Agents and Multi-Agent Systems Conference, Sydney, Australia, 2003.

Wang, D., & Narayanan, S. (2002). *A confidence-score based unsupervised MAP adaptation for speech recognition.* Paper presented at the Proceedings of 36th Asilomar Conference on Signals, Systems and Computers.